情報処理学会論文誌 Vol. 50 No. 4 1375–1389 (Apr. 2009)

# 複数モデルの統合によるLDAトピックモデルの高精度化とテキスト入力支援への応用

中村 明[†1] 速水 悟[†2] 津田裕亮[†2]
松本忠博[†2] 池田尚志[†2]

単語間の大域的な依存関係をトピック（話題）としてモデル化する言語モデルの1つであるLDA（Latent Dirichlet Allocation）を複数個統合する方式によって，言語モデルを高精度化・安定化できることを示す．新聞記事コーパスを用いた実験の結果，提案方式では単一のLDAからなる同一規模のモデルと比較して，つねに推定精度が向上・安定化することを確認した．単一LDAでは潜在トピック数 $C = 100$ 前後を境に性能が低下するのに対し，提案方式では過適応が抑制され，はるかに大きい総トピック数（＝各モデルの潜在トピック数の総和）まで性能が向上し続ける．また提案方式によるunigram確率を用いて $N$-gram確率（$N \geqq 2$）を補間することにより，trigramのパープレキシティを従来方式より大幅に削減できる．さらに本論文では，提案方式を予測入力に基づくテキスト入力支援（predictive text entry）に応用することを想定し，テキスト入力支援に適した言語モデル評価指標i-PPを提案する．この指標はパープレキシティの拡張であり，任意文字数の読み入力時点における平均単語分岐数を表す．この指標を用いた評価の結果，提案手法では入力読み文字数 $l = 2$ の時点まで通常のパープレキシティと同程度にi-PPを削減でき，従来方式よりも高精度に予測候補を絞り込めることが確かめられた．

# Integration of Multiple LDA Topic Models and Its Application to Predictive Text Entry

Akira Nakamura,[†1] Satoru Hayamizu,[†2]
Yusuke Tsuda,[†2] Tadahiro Matsumoto[†2]
and Takashi Ikeda[†2]

This paper describes a method that improves accuracy and stability of a language model. The method integrates multiple units of LDA (Latent Dirichlet Allocation), which is a probabilistic language model that models long-range dependencies among words as topics. The experiment on news text corpora shows that the proposed method constantly makes its performance more precise and stable, comparing to the single LDA with almost the same number of the model parameters. The perplexity of the method remains decreasing until the total number of latent topics reaches far larger than that of single LDA, whereas the perplexity of the single LDA turns to increase due to overfitting as the number of latent topics is around 100. In particular, the proposed method significantly reduces the trigram perplexity by combining with the topic-dependent unigram probabilities. This paper also proposes a new evaluation measure i-PP suitable for evaluating a language model applied to predictive text entry. This measure, which is an extended perplexity, indicates the average number of word choices when any length of phonetic (hiragana) substring is input. Evaluation with this measure demonstrates that the proposed method decreases i-PP by the same rate as the common perplexity until the substring length $l = 2$, reducing candidates with higher accuracy than the existing method.

## 1. はじめに

単語間の局所的な依存関係をモデル化する $N$-gram モデル[1] は，シンプルであるがゆえに汎用性が高く，様々な自然言語処理タスクに応用されている．$N$-gram モデルに単語間の大域的な依存関係を組み込むことによって，言語をより精緻にモデル化することが可能であり，キャッシュモデルやトリガモデルの併用はその一例である[2),3)]．

キャッシュモデルやトリガモデルが単語間の長距離の依存関係を単語（対）の形で直接，モデル化するのに対し，近年，単語間の大域的な依存関係を話題（トピック）としてモデル化するトピックモデルの研究が進展している．unigram の混合モデルである Mixture of Unigrams[4]，潜在トピックを導入し LSI（Latent Semantic Indexing）を確率モデルとして再定式化した PLSI（Probabilistic Latent Semantic Indexing）[5]，PLSI をベイズ学習に基づき改良した LDA（Latent Dirichlet Allocation）[6]，単語生起確率を混合ディリクレ分布に従う確率変数とする DM（Dirichlet Mixture）[7] 等が提案されている．

これらトピックモデルでは文脈に基づいて unigram 確率を動的に推定することによりパープレキシティを削減できる．そして unigram rescaling[8] 等の補間手法によって $N$-gram 確率（$N \geqq 2$）をトピックに適応させることが可能であり，連続音声認識，同音異義語のかな

†1 三洋電機株式会社エコロジー技術研究所
ECO Technology Research Center, SANYO Electric Co., Ltd.
†2 岐阜大学工学部応用情報学科
Department of Information Science, Gifu University

漢字変換誤り検出等への適用が試みられている[9)–11)]．また Blei らは LDA を文書分類のための特徴抽出器としても用いている[6)]．筆者らの一部は以前，$N$-gram モデルとキャッシュモデルを用いた予測入力によるテキスト入力支援（predictive text entry）システムを構築し，医療文書の入力支援における有効性を確認しているが[12)]，トピックモデルを用いることによってこれをさらに高精度化することを目指している．

言語モデルに限らず，有限個の事例に基づいてパラメータを推定する学習器では，一般にモデルのパラメータ数増加にともなって open data に対する精度（汎化能力）が低下する，いわゆる過適応が不可避である．$N$-gram におけるスパースネスの問題もこの一種であり，PLSI や LDA 等のトピックモデルにおいても潜在トピック数の増加にともない過適応が問題となる．そのため，計算コストを考慮しつつ最適なトピック数を決定する必要がある．

トピックモデルにおいて過適応の問題に対処した先行研究としては，DM（Dirichlet Mixture）のパラメータ推定において階層ベイズモデルを用いて平滑化を行った報告がある[13)]．文献 7) では，モデル規模（混合ディレクレ分布の混合数）の異なる複数の DM を重み付きで平均することにより過適応を抑制できることが示されている．また，階層ディリクレ過程を用いて LDA の最適な潜在トピック数を決定する手法も提案されている[14)]．

これに対し本論文では，独立に学習した複数個の LDA モデルによって得られた推定結果を集団学習の枠組みで統合する方式[15)] を提案し，単一の LDA による構成と比較して同程度のモデル規模（= 潜在トピック数 × モデル数）で高精度かつ安定した性能が実現できることを示す．そしてこの方式をテキスト入力支援に応用することを想定し，提案方式を用いて $N$-gram をトピック適応させた場合の推定精度と安定性を検証する．さらにテキスト入力支援に適した言語モデル評価指標として，テストセットパープレキシティを拡張した新たな評価指標を提案し，この指標を用いた評価を行う．

以下，2 章で LDA の概要について述べ，3 章で提案するシステムの構成と言語モデル評価指標について説明する．そして 4 章で評価実験の手順，5 章で実験結果を示し，6 章で考察を行う．

## 2. LDA の概要

LDA（Latent Dirichlet Allocation）[6)] は，各潜在トピック $(z_1, z_2, \ldots, z_C)$（$C$：潜在トピック数）の生成確率 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_C)$ が多項分布の共役事前分布であるディリクレ分布 $\mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha})$ に従うと仮定した文書生成モデルである．文書 $d = (w_1, w_2, \ldots, w_{|d|})$ の出現確率は次式で表される（$|d|$ は文書 $d$ の総単語数を表す）．

$$p(d \,|\, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) \left( \prod_{n=1}^{|d|} \sum_{k=1}^{C} p(w_n \,|\, z_k, \boldsymbol{\beta}) p(z_k \,|\, \boldsymbol{\theta}) \right) d\boldsymbol{\theta} \tag{1}$$

$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ が LDA のモデルパラメータであり，$\beta_{kj}$ はトピック $z_k$ における語 $w_j$ の unigram 確率 $p(w_j \,|\, z_k)$ を表す（$1 \leqq j \leqq V$）（$V$：語彙数）．$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_C)$ はディリクレ分布

$$\mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{C} \alpha_k\right)}{\prod_{k=1}^{C} \Gamma(\alpha_k)} \prod_{k=1}^{C} {\theta_k}^{\alpha_k - 1} \tag{2}$$

のパラメータである．パラメータ $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ の学習には変分ベイズ法[6)] による近似計算や MCMC（Markov Chain Monte Carlo）[16)] が用いられるが，本論文では変分ベイズ法を用いる．

未知の文脈 $h$ に対する文脈適応は，学習時と同様の変分近似により計算する．すなわち，$h$ に対する変分パラメータ $\gamma_k$ および $\phi_{kj}$ を導入し，学習済みの $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ を用いて以下の手順を収束するまで繰り返す．

$$\mathrm{VB-Estep:}\ \phi_{kj} \propto \beta_{kj} \exp(\Psi(\gamma_k)) \tag{3}$$

$$\mathrm{VB-Mstep:}\ \gamma_k = \alpha_k + \sum_{j=1}^{V} n(h, w_j) \phi_{kj} \tag{4}$$

$\Psi(\gamma)$ は digamma 関数であり，$n(h, w_j)$ は $h$ における語 $w_j$ の出現回数を表す．得られた $\gamma_k$ を文脈 $h$ のもとでの各潜在トピックの混合比とする．したがって，文脈 $h$ のもとでの語 $w_{j'}$ の生起確率は次式により与えられる．

$$p(w_{j'} \,|\, h) = \frac{\sum_{k=1}^{C} \gamma_k \beta_{kj'}}{\sum_{k=1}^{C} \gamma_k} \tag{5}$$

LDA はトピックの事前分布にディリクレ分布を用いることにより，トピックの広がりやトピック間の関係を表現できる点で PLSI より優れている．またベイズ推定に基づくため過適応の問題が少ないとされている．

## 3. 提案システム

### 3.1 複数 LDA の統合

本論文で提案する言語モデルの構成を**図 1** に示す．本システムは独立に学習した $M$（$\geqq 2$）個の LDA モデル $Q_1, Q_2, \ldots, Q_M$ を持つ．各モデルの潜在トピック数は必ずしも同じでな

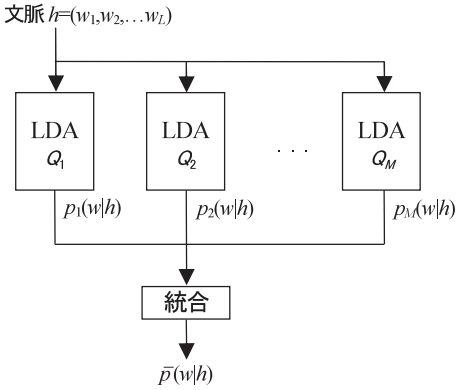



**図 1**　複数 LDA 統合型言語モデル
Fig. 1　The language model integrating multiple LDA's.

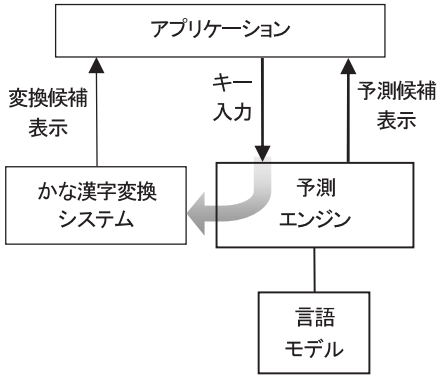



**図 2**　テキスト入力支援システム
Fig. 2　Predictive text entry system.

くてもよいが，本論文では主に各モデルの潜在トピック数が同じ場合を扱う．以降，複数個の LDA からなる図 1 の構成を **m-LDA**（multiple-LDA）と呼ぶ．

$L$ 形態素からなる文脈 $h$ が入力されると，各 LDA モデル $(Q_1, Q_2, \ldots, Q_M)$ はそれぞれトピック推定を行い，式 (5) により $h$ のもとでの語 $w$ の unigram 確率 $p_m(w\,|\,h)$ を求める $(1 \leqq m \leqq M)$．そして $M$ 個の $p_m(w\,|\,h)$ の平均値

$$\bar{p}(w\,|\,h) = \frac{1}{M}\sum_{m=1}^{M} p_m(w\,|\,h) \tag{6}$$

を最終的な語 $w$ の unigram 確率の推定値として出力する．$p_m(w\,|\,h)$ を適当に重み付けして平均する方式も考えられるが本論文では扱わない．以下，本論文では式 (6) による unigram 確率をトピック依存 unigram 確率と呼ぶ．

### 3.2　テキスト入力支援システム

前節で示した言語モデルを用いたテキスト入力支援システムの構成を**図 2** に示す．本システムは Windows OS 上で MS-IME や ATOK 等既存のかな漢字変換システムとともに動作する．本システムはユーザのキー入力をつねに監視しており，かな漢字変換が ON の場合に限り，キー入力のたびに入力文字列を予測エンジンに送る．予測エンジンは，読み[⋆1]が入力文字列に前方一致する語の中から，言語モデルに基づいて予測候補単語のリストを生成し画面に表示する．

ユーザは予測候補リスト中に入力したい単語があれば上下矢印キー等により選択し，なければ次の読み文字を入力する．予測候補を選択したら，予測エンジンはさらに後続の語を予測し予測候補リストを表示する．

予測候補リストは 1 文字追加入力されるたびに更新される．すべての読みを入力しても候補リスト中に入力したい単語が現れない場合には，変換キー（スペースキー等）の押下によりかな漢字変換候補が表示され，通常のかな漢字変換による入力操作に移行する．この際，予測候補リストは非表示となる．すなわち，本入力支援システムは既存のかな漢字変換システムとともに協調して動作する．

**図 3** に本システムにおけるテキスト入力過程の例を示す．同図 (a) は「昨年 8 月、当院にて」に続いて読み「じょう」を入力した状態であり，既入力単語列につながり読みが「じょう」で始まる予測候補のリストが表示されている．同図 (b) は，(a) において入力したい語が候補リスト中にないために次の読み文字「ぶ」を入力したところ，入力しようとする語「上部消化管」が予測候補に現れたため，これを選択した状態を表している．同図 (c) は「上部消化管」を選択確定後さらに後続の予測候補が表示され，「内視鏡検査」を選択した状態である．このようにシステムが精度良く候補を予測できれば，効率良くテキスト入力を行うことができる．

⋆1 英語等表記を直接入力する言語では，表記が入力文字列に前方一致する語が予測候補となる．

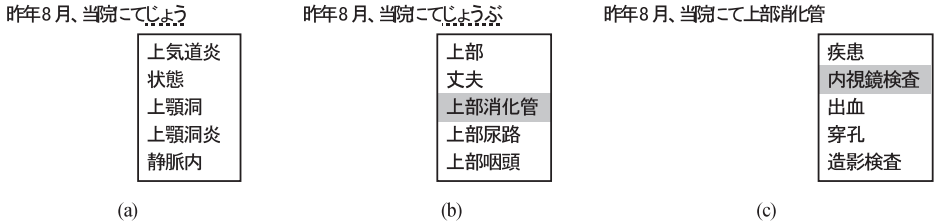



**図 3**　テキスト入力過程の例
Fig. 3　An example of the text entry process.

なお，本論文では言語モデルとして複数 LDA からなる図 1 のトピックモデルを用いるが，図 2 のテキスト入力支援システムは特定の言語モデルを前提としない一般的な枠組みであり，任意の言語モデルと組み合わせることが可能である．

### 3.3　テキスト入力支援のための評価指標

言語モデルの評価指標としては，一般に次式のテストセットパープレキシティ（TPP; 以降，単に PP と呼ぶ）が用いられる．

$$PP(D\,|\,Q) = 2^{H(D|Q)} \tag{7}$$

$Q$ は言語モデル，$D$ は $L_D$ 個の形態素 $w_1^{L_D} = w_1 \ldots w_{L_D}$ からなる評価テキスト，$H(D\,|\,Q)$ は言語モデル $Q$ による評価テキスト $D$ に対するクロスエントロピーであり $H(D\,|\,Q) = -(1/L_D)\log p_Q(w_1^{L_D})$ である（$p_Q(\cdot)$ は言語モデル $Q$ による単語（列）の生成確率を表す）．

PP は言語モデルで評価テキストを予測した場合における単語の平均分岐数を表し，テキスト入力支援においてもモデルの基本的な評価指標として有用である．ただしテキスト入力支援では，ユーザが読みを入力するたびに候補が絞り込まれる．したがって「読みを何文字入力すれば候補リスト内に正解単語が含まれると期待できるか？」という観点からモデルを評価できることが望ましい．

そこで本論文では，テキスト入力支援に適した言語モデル評価指標として任意の入力文字数における平均単語分岐数を表す i-PP（incremental/interactive-PP)[⋆1]を提案する．入力読み文字数 $= l$ における i-PP は次式で定義される．

$$iPP(l) = 2^{H(D|Q,l)} \tag{8}$$

$$where$$

$$H(D\,|\,Q,l) = -\frac{1}{L_D}\sum_{n=1}^{L_D}\log p(w_n\,|\,w_1^{n-1}, s_n(1:l))$$

ただし $s_n(i:j)$ は語 $w_n$ の読み文字列 $s_n$ の $i$ 文字目〜$j$ 文字目からなる部分列を表す．$p(w_n\,|\,w_1^{n-1}, s)$ は入力単語列 $w_1^{n-1}$ および入力読み文字列 $s$ のもとでの語 $w_n$ の生起確率を表し，$w_n$ の読みが $s$ に前方一致する場合のみ $p(w_n\,|\,w_1^{n-1}, s) > 0$ となるよう次式により計算する．

$$p(w_n\,|\,w_1^{n-1}, s) = \begin{cases} p(w_n\,|\,w_1^{n-1}) \Big/ \displaystyle\sum_{w\in V^{(s)}} p(w\,|\,w_1^{n-1}) & w \in V^{(s)} \\ 0 & w \notin V^{(s)} \end{cases} \tag{9}$$

$V^{(s)}$：読みが文字列 $s$ に前方一致する語の集合

入力読み文字数 $= 0$ の場合の i-PP，すなわち $iPP(0)$ は通常の PP に一致する．したがって i-PP は PP の自然な拡張となっている．

## 4. 評 価 実 験

### 4.1　学習データおよび評価データ

学習用データおよび評価用データを以下に示す．

［学習用データセット］

CD 一毎日新聞 2005 データ集[17] 全記事（95,881 件）．約 2,864 万形態素，異なり語数 185,196

［評価用データセット］

CD 一毎日新聞 2006 データ集[17] のうち，200 文字以上の記事から無作為抽出した 1,000 件，約 40 万形態素

学習用データセット・評価用データセットとも，文節構造解析システム ibukiC[18] により形態素解析を行った．評価データ中，学習データに含まれない未知語はのべ 1,746 語であった．

⋆1 入力のたびに単語を絞り込み interactive に候補を表示する方式を incremental search と呼ぶことから命名している．

### 4.2 学習条件

潜在トピック数 $C = 20$，30，50，70 の場合に対し，それぞれ以下に示す 2 通りの方法で LDA の学習を行った．

(1) **m-LDA1**（同一の学習データセット使用）：前節で示した学習用データ 95,881 件（以降，学習データセット $\boldsymbol{T}_0$ と呼ぶ）を学習データセットとし，$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ に異なる初期値を与えて $M$ 回の学習を行い，$M$ 個の LDA を構築する．

(2) **m-LDA2**（異なる学習データセット使用）：$\boldsymbol{T}_0$ から復元抽出により 95,881 記事を抽出する作業を $M$ 回行って $M$ 個のデータセット $\boldsymbol{T}_{b1}, \boldsymbol{T}_{b2}, \ldots, \boldsymbol{T}_{bM}$ を作成，これらを学習データセットとして $M$ 回の学習を行い，$M$ 個の LDA を構築する．

なお (2) で用いる $M$ 個のデータセット $\boldsymbol{T}_{b1}, \boldsymbol{T}_{b2}, \ldots, \boldsymbol{T}_{bM}$ の組はすべての潜在トピック数に対して共通のものを用いた．データセット $\boldsymbol{T}_{b1}, \boldsymbol{T}_{b2}, \ldots, \boldsymbol{T}_{bM}$ 各々の異なり記事数は平均 60,743 記事であった．

上記 (1)，(2) とも，元の学習データセット $\boldsymbol{T}_0$ における出現回数が 4 回以下の語を除いた 77,794 語で学習を行った．パラメータの初期化は，(1)，(2) とも $\boldsymbol{\alpha}$ については区間 $[0, 1]$ の一様乱数 $+1$ で初期化し，$\boldsymbol{\beta}$ については潜在トピックごとに $V$（$= 77{,}794$）個の一様乱数を和が 1 になるよう正規化した値を初期値とした．学習アルゴリズムは 2 章で述べた変分ベイズ法を使用し，パラメータ $\boldsymbol{\alpha}$ の推定には Fixed-point iteration [19] を用いた．収束判定は，学習データセットに対する 1 ステップ前からのパープレキシティの減少幅が 0.1%未満となった時点で収束とした．

LDA の学習とは別に，学習データセット $\boldsymbol{T}_0$ を用いてトピック非依存の $N$-gram モデル（$N \leqq 3$）を構築した．bigram および trigram 確率の推定にはカッツ・スムージング [20] による平滑化を用いた．これらのトピック非依存 $N$-gram 確率は後述の unigram rescaling の際に用いられる．

### 4.3 評価方法

4.1 節に示した評価用データセットに対して提案システムにより $N$-gram 確率（$N = 1$，2，3）を求め，PP および i-PP により評価する．トピック依存 $N$-gram 確率（$N = 2$，3）は，式 (6) により求めたトピック依存 unigram 確率 $\bar{p}(w_n \,|\, h)$ から次式の unigram rescaling [8] により算出する．

$$p(w_n \,|\, w_{n-N+1}^{n-1}, h) \propto \frac{\bar{p}(w_n \,|\, h)}{p(w_n)}\, p(w_n \,|\, w_{n-N+1}^{n-1}) \tag{10}$$

$p(w_n)$：トピック非依存 unigram 確率

$p(w_n \,|\, w_{n-N+1}^{n-1})$：トピック非依存 $N$-gram 確率

LDA に与える文脈 $h$ の長さは予備実験により 20 形態素とした．評価用データセット中の各記事の先頭 20 形態素まではトピック推定を行わず，トピック非依存の $N$-gram 確率をそのまま用いる．また，LDA 学習時の語彙に含まれない低頻度語については，つねにトピック非依存 $N$-gram 確率を用いる．

## 5. 実験結果

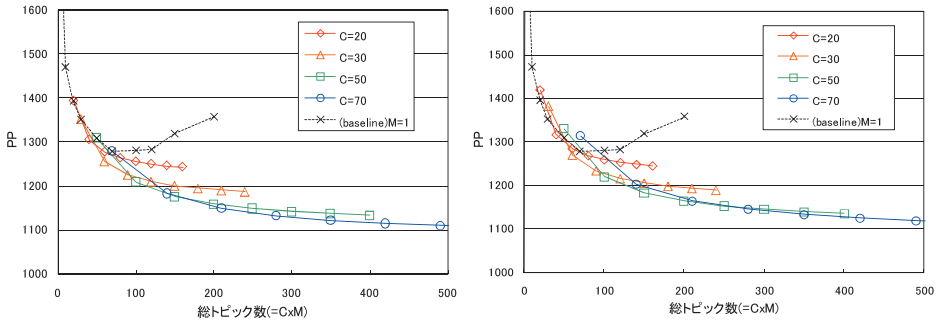
### 5.1 総トピック数と学習条件の違いによる比較

前章で示した学習用データセットおよび評価用データセットを用いて，各 LDA の潜在トピック数 $C = 20$，30，50，70 の場合について，unigram の PP によりモデル数 $M = 8$ までの評価を行った結果を**図 4** に示す．システム全体のモデル規模が同等の条件下で比較するため，横軸は総トピック数（= 潜在トピック数 $C \times$ モデル数 $M$）とした．$C = 20$，30，50，70 の各プロットは，8 個のモデルから $M$ 個（$1 \leqq M \leqq 8$）を選ぶすべての組合せ（${}_8\mathrm{C}_M$ 通り）についての平均値である．baseline はデータセット $\boldsymbol{T}_0$ で学習した単一の LDA（すなわち $M = 1$）で $C$ を変化させた場合の性能を示す．なお，グラフの範囲外のため図示されていないがトピック非依存の unigram モデルによるパープレキシティ（$PP_0$）は 1954.9 であった．

図 4 に示すように，baseline では $C = 100$ 前後を境に性能が悪化するのに対し，提案手法では **m-LDA1**，**m-LDA2** ともに総トピック数の増加にともない PP が単調に減少し，baseline に比べて PP を最大 13%削減できている．$M \geqq 2$ ではすべての場合において baseline の性能を上回っており，複数 LDA の統合によって同規模の（すなわち $C \times M$ が等しい）単一 LDA よりつねに性能が改善されることが分かる．**m-LDA1** と **m-LDA2** とを比較すると，**m-LDA2** のほうが統合による PP の減少がやや大きい．ただし統合前（$M = 1$）の性能が **m-LDA1** に比べやや劣るため，統合後の性能はほぼ同程度である．

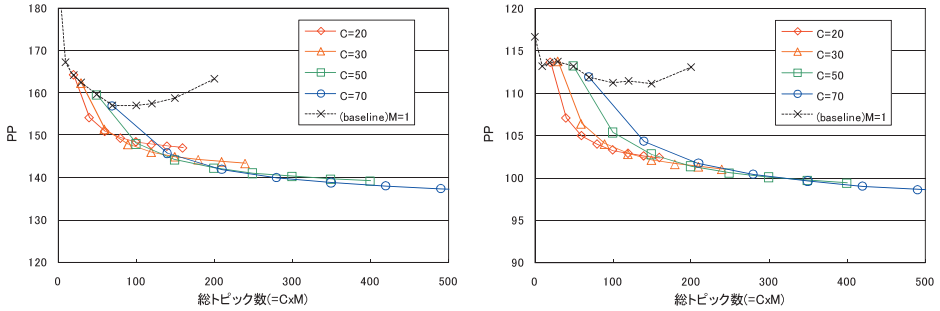
### 5.2 トピック依存 $N$-gram による評価

トピック依存 $N$-gram 確率（$N = 2$，3）により提案手法（**m-LDA1** のみ）の評価を行った結果を**図 5** に示す．unigram の場合と同様，提案手法では総トピック数の増加とともに PP が減少している．特に trigram の場合，baseline ではトピック非依存 trigram による PP（$= 116.7$）からの減少が最大でも 4.6%（116.7 → 111.3）にすぎないのに対し，提案手法では最大 15.5%（116.7 → 98.6）削減できている．この結果は，単一 LDA では $N$-gram の




(a) unigram (m-LDA1)　(b) unigram (m-LDA2)

**図 4**　総トピック数の違いによる PP の比較
Fig. 4　Perplexity with various total numbers of latent topics.




(a) bigram (m-LDA1)　(b) trigram (m-LDA1)

**図 5**　bigram および trigram の PP
Fig. 5　Perplexity of bigram and trigram.

PP を減らすことが困難であるという問題を提案手法により改善できることを示している．

提案手法により $N$-gram の推定精度が向上する事例を**図 6** に示す．同図 (a)〜(j) は，$C = 50$ の 8 個のモデルのうち 1 番目のモデル単独で予測を行った場合に対して，1〜4 番目のモデルを統合した場合の trigram の PP 削減率が高い上位 10 記事の見出しを示している．末尾の数字はモデル 1 単独の場合に対する 4 モデル統合時の trigram PP の相対値を表す．

教育や学校に関連する内容の記事が上位 10 記事の中に多く含まれている（(a)，(b)，(d)，(f)，(h)）．詳細は次章で考察するが，これは単一の LDA ではある話題に対して適切にトピックを推定できない事態が生じるが，提案手法によってこれを改善できることを表している．

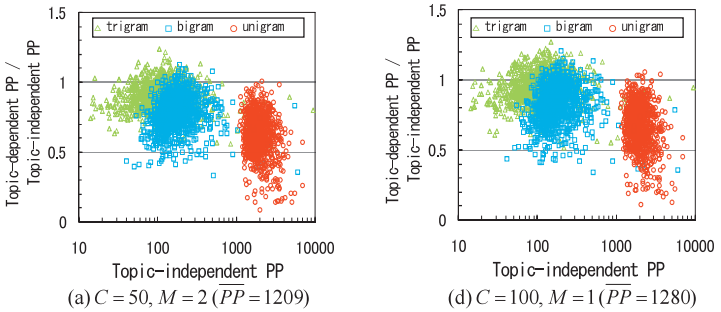
(a) 学生音コン：全国大会　バイオリン部門　小学校は岡本さん、中学校は成田さんが1位 [0.710]
(b) 高校履修不足：単位取得「7日間の旅」 埼玉の私立高、豪修学旅行参加で認定 【大阪】[0.723]
(c) 捜査資料流出：ウィニーの使用、緊急点検を指示ーー警察庁 [0.733]
(d) 大学入試：国公立大2次試験出願状況　3日午前10時現在（その1）（7日午後3時現在　文部科学省集計）[0.735]
(e) 自衛隊イラク派遣：第10次群が近く出国へ [0.736]
(f) 大学入試：国公立大2次試験出願状況（その3） [0.740]
(g) サッカー：J1　ガ大阪2ー0甲府　ガンバ、首位返り咲き（第12節・6日） [0.754]
(h) 現代的教育ニーズ取組支援プログラム：最多112件選定 [0.756]
(i) 対馬暖流：秋の流量、冬の降水量と比例　九大など研究、積雪予測へ一歩 [0.757]
(j) 停電：鳥の巣原因？2700世帯でーー兵庫・尼崎 【大阪】[0.757]

**図 6**　提案手法で PP が改善した記事の例
Fig. 6　Examples of articles of which the PP was improved by the proposed method.

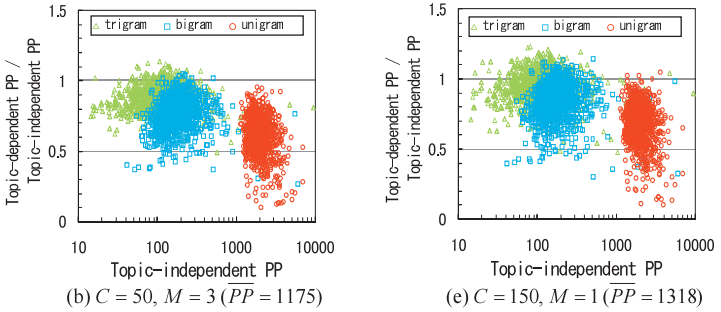
### 5.3　モデル出力の安定性評価

システム構成の違いによる評価用データセット中の各記事に対する PP の傾向を比較した結果を**図 7** と**図 8** に示す．評価用データセットは 4.1 節で示した 1,000 記事である．図 7 の縦軸は，それぞれの構成におけるトピック依存 $N$-gram（$N = 1$，2，3）による各評価記事に対する PP をトピック非依存 $N$-gram による PP を基準とした相対値で表しており，値が 1.0 より小さければトピック非依存 $N$-gram より推定精度が改善されていることを意味する．図 8 の縦軸も同様である．両図とも横軸はトピック非依存 $N$-gram による各記事に対する PP であり，$\overline{PP}$ はトピック依存 unigram による全評価記事に対する PP を示す．

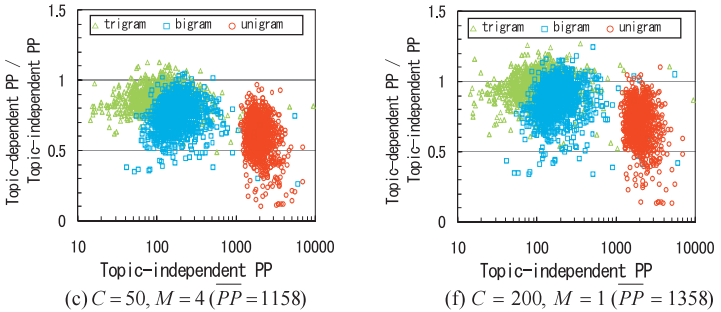
図 7 (a)〜(c) は **m-LDA1** で $C = 50$ に固定し $M$ を 2，3，4 とした場合，(d)〜(f) は $M = 1$ とし（単一 LDA）$C = 100$，150，200 と変化させた場合を表す．図より **m-LDA1**（(a)〜(c)）ではモデル数の増加にともなってプロットが下方に移動し推定精度が向上している．$M = 2$，3，4 における PP 相対値の平均を比較すると，unigram で 0.680，0.627，




(a) $C = 50$, $M = 2$ ($\overline{PP} = 1209$)　(d) $C = 100$, $M = 1$ ($\overline{PP} = 1280$)



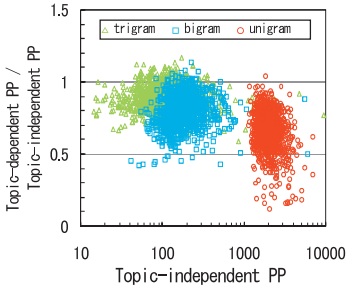

(b) $C = 50$, $M = 3$ ($\overline{PP} = 1175$)　(e) $C = 150$, $M = 1$ ($\overline{PP} = 1318$)



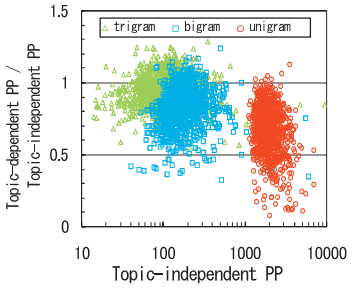

(c) $C = 50$, $M = 4$ ($\overline{PP} = 1158$)　(f) $C = 200$, $M = 1$ ($\overline{PP} = 1358$)

**図 7**　各記事に対する PP（1）—総トピック数が等しいモデル間の比較—

Fig. 7　Perplexity for each article (1) — Comparison between the models with the same total number of latent topics.




(a) $C = 20$, $M = 3$ ($\overline{PP} = 1277$)




(b) $C = 70$, $M = 1$ ($\overline{PP} = 1278$)

**図 8**　各記事に対する PP（2）—平均推定精度が等しいモデル間の比較—

Fig. 8　Perplexity for each article (2) — Comparison between the models with the same accuracy.

0.599，trigram で 0.972，0.905，0.869 と改善されている．分散も同様に unigram で 0.167，0.150，0.140，trigram で 0.114，0.096，0.086 と改善されており，モデル統合によって性能が向上・安定化する．一方，単一モデルのままトピック数を増やした場合（(d)〜(f)）は推定精度が向上しない．$C = 100$，200 における PP 相対値の平均は unigram で 0.666，0.698，trigram で 0.954，0.969 となり，$C = 50$，$M = 1$ の場合（unigram で 0.680，trigram で 0.972；前述）からほとんど改善されていない．総トピック数 $C \times M$ が同一の **m-LDA1** と **m-LDA2** とを比較すると，精度が向上する記事と低下する記事がともに増え，文脈によって当たり外れの大きい不安定な特性を示すようになる．

図 8 は $\overline{PP}$ がほぼ等しい構成での記事ごとの PP の違いを示す．$C = 20$，$M = 3$ と $C = 70$，$M = 1$ とでは全記事に対する平均推定精度は同等であるにもかかわらず，$C = 70$，$M = 1$ の場合は記事による精度のバラツキが大きく，トピック非依存 $N$-gram より悪化

**表 1**　トピック非依存 $N$-gram の i-PP
Table 1　i-PP of topic-independent $N$-gram.

| | 入力読み文字数 $l$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| unigram | 1954.9 | 32.40 | 14.52 | 3.10 | 1.75 | 1.46 |
| bigram | 188.9 | 9.27 | 6.26 | 2.31 | 1.57 | 1.38 |
| trigram | 116.7 | 7.66 | 5.65 | 2.27 | 1.58 | 1.39 |

するケースも少なくない．対して $C = 20$，$M = 3$ では安定して精度が向上しており，提案方式ではより少ないモデル規模で同等精度かつ安定したモデルが得られることを示している．

### 5.4　テキスト入力支援を想定した評価

3.3 節で提案したテキスト入力支援のための言語モデル評価指標 i-PP を用いて，入力読み文字数 $l = 0$〜5 の場合にトピック非依存 $N$-gram モデルを評価した結果を**表 1** に示す（$l = 0$ における値は通常の PP に等しい）．評価用データセットは 4.1 節で示した 1,000 記事である．

読みが入力されるとパープレキシティは急速に減少し，unigram/bigram/trigram ともに $l = 4$ で 2 未満となる．これは次単語を予測する場合，読みを 4 文字入力すれば unigram でも平均 2 単語以内に絞り込めることを表している．一方，$l = 0$〜2 の範囲では bigram および trigram と unigram との差は大きく，bigram または trigram を用いれば読み入力文字数が 1〜2 文字でも 10 単語以内に絞り込める可能性が高い．ユーザに提示する候補単語数は多くても 10 語以下，できれば 5 語程度であることが望ましいため[*1]，トピックモデルを用いて予測することにより $l = 1$〜2 のときの i-PP を 5 以下に抑えることができれば，テキスト入力支援において有効に機能することが期待できる．

提案方式である **m-LDA1** および単一 LDA を用いて bigram および trigram をトピック適応させた場合の，入力読み文字数 $l = 1$〜5 における i-PP を**図 9** に示す．$l = 1$〜2 ではトピックモデルを用いることにより bigram，trigram ともにパープレキシティを削減できており，単一 LDA（$C = 70$）よりも **m-LDA1**（$C = 70$，$M = 4$）のほうが削減幅が大きい．$l \geqq 3$ ではトピック非依存 $N$-gram との差はわずかとなる．**m-LDA1**（$C = 70$，$M = 4$）

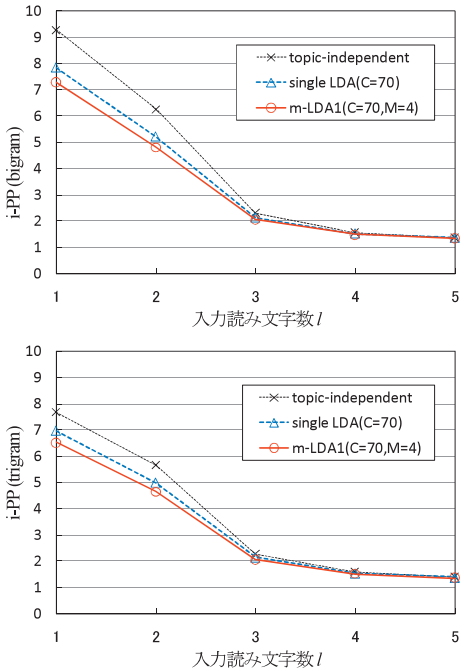



**図 9**　トピックモデルによる i-PP
Fig. 9　i-PP obtained by topic models.

では $l = 2$ のとき bigram で 23.3%（6.26 → 4.80），trigram で 17.9%（5.65 → 4.64）i-PP が削減され i-PP の値が 5 未満となっている．これは提案方式をテキスト入力支援に応用することにより，読みを 2 文字入力すれば平均 5 単語以内に予測候補を絞り込めると期待できることを意味する．$C = 70$，$M = 4$ の場合におけるトピック依存 $N$-gram による i-PP の値を表 1 と対比して**表 2** に示す．

トピック非依存 $N$-gram による i-PP を 1.0 とした場合のトピック依存 $N$-gram による i-PP の相対値を**図 10** に示す．unigram/bigram/trigram とも $l = 1$〜2 では $l = 0$ のとき（すなわち通常の PP）と同程度の削減効果が得られている．$C = 70$，$M = 4$ の構成の場合，削減率は unigram で約 40%，bigram で約 25%，trigram で約 15%である．

⋆1 人間が 1 度に処理できる項目数は 7 ± 2 個といわれている[21]．

**表 2**　トピック依存 $N$-gram の i-PP（$C = 70$，$M = 4$）
Table 2　i-PP of topic-dependent $N$-gram ($C = 70$, $M = 4$).

| | 入力読み文字数 $l$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| unigram | 1164.4 | 20.39 | 8.63 | 2.44 | 1.57 | 1.38 |
| bigram | 142.6 | 7.28 | 4.80 | 2.04 | 1.48 | 1.34 |
| trigram | 102.2 | 6.52 | 4.64 | 2.05 | 1.50 | 1.35 |

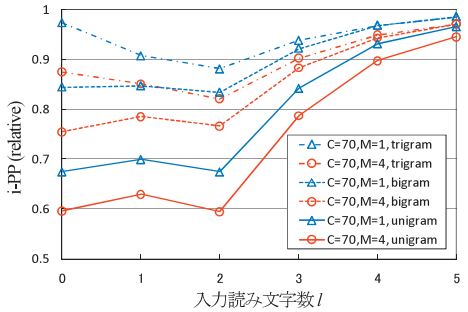



**図 10**　トピック非依存 $N$-gram に対する i-PP の相対値
Fig. 10　i-PP relative to topic-independent $N$-gram.

## 6. 考　　察

本章では，前章の実験結果をふまえ提案方式の有効性について，学習条件との関係，改善事例の分析，先行研究との比較，テキスト入力支援に応用した際の実用性の側面から考察する．

### 6.1　提案方式の効果と学習条件との関係

前章 5.1〜5.3 節では，独立に学習した複数の LDA による推定結果を統合する提案方式によって $N$-gram を高精度化・安定化できることを示した．ここでは提案方式の効果と学習条件との関係について考察し，さらに提案方式を実装する際に最適な構成（潜在トピック数 $C$ およびモデル数 $M$）を決めるための指針を示す．

（1）統合モデル数 $M$ と統合効果との関係

**図 11** は 5.1 節で示した実験結果における統合モデル数 $M$ と PP との関係を表しており，

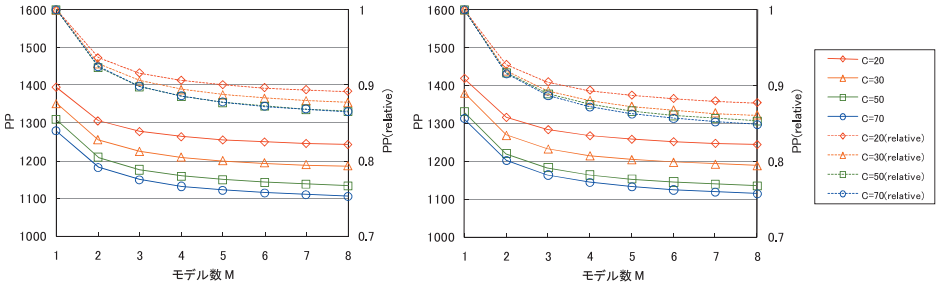



**図 11**　統合モデル数の違いによる PP の比較
Fig. 11　Perplexity with various numbers of the integrated models.

1 モデルあたりの潜在トピック数 $C = 20$，30，50，70 それぞれについて $M = 1$（単一 LDA）の場合を 1.0 とした PP の相対値（縦軸右）をあわせて示してある．

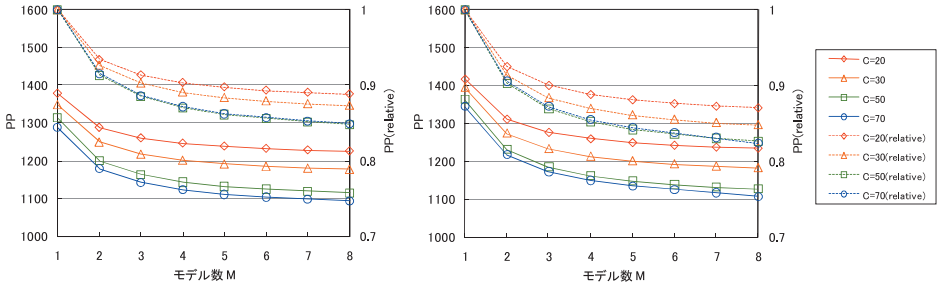
図より，**m-LDA2** のほうが統合による PP 削減効果は大きいものの，$M = 1$ のときの精度がやや劣るため統合後の PP は同程度かそれ以下となっている．複数の学習器を統合する場合，一般には個々の学習器が異なった傾向を持つほど効果が大きい．**m-LDA2** が **m-LDA1** に比べ統合による効果が大きいのは，復元抽出により生成した異なる学習データを用いたことによってモデル間の差異が拡大したためである．

今回は $M = 8$ までの範囲で評価を行ったが，特に **m-LDA2** では $M = 8$ でもまだ統合効果が飽和していないため，$M$ を増やすことによりさらに高精度化できる可能性がある．

（2）学習データセットサイズとの関係

**図 12** は筆者らが先に報告した実験結果[15] について，図 11 と同様モデル数 $M$ と PP との関係を示したものである．この実験では本論文での学習用データセットの約半分（毎日新聞 2005 データ集のうち 48,035 記事）を学習データとして用いた．LDA 学習時の語彙数は 75,314 語で，本論文の実験とほぼ同程度である．

図 11 と図 12 とを比較すると，**m-LDA1**，**m-LDA2** とも図 12 のほうが図 11 より統合効果が大きく，統合後の精度は同程度となっている．学習コーパスのサイズが小さい場合，個々のモデルはより過適応しやすくなるため単独での精度は低下するが，このときにモデル間の差異が拡大するために統合効果が大きくなったものと考えられる．これにより提案方式では約半分の学習データで同程度の精度が得られている．この結果は，学習データ量が不十分なために単一 LDA では高い精度を得ることが難しい場合に本提案手法によって精度を向




(a) m-LDA1　(b) m-LDA2

**図 12**　統合モデル数の違いによる PP の比較（2）—学習データのサイズが半分の場合—

Fig. 12　Perplexity with various numbers of the integrated models (2) — learned with half of the learning data set.

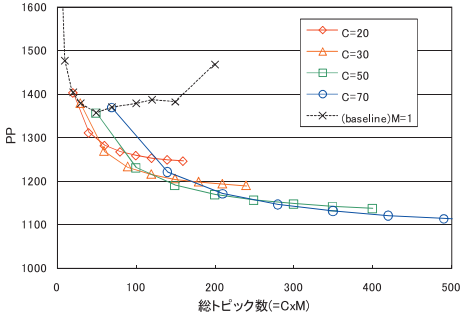



**図 13**　$\alpha$ のみ乱数初期化時における提案手法の効果

Fig. 13　The effect of the proposed method when only $\alpha$'s are initialized with random numbers.

上させることができることを示している．

（3）学習データセットの構成方法との関係

4.2 節で述べたように，**m-LDA2** の学習に用いたデータセットの異なり記事数は平均 60,743 記事であった．これは全 95,881 記事の約 37%は各データセットに含まれていないことを意味する．抽出される記事はデータセットごとに異なるため，モデル間の差異が拡大しさらに高精度化できることを期待した．しかし実験の結果，図 11，図 12 ともに統合による PP 削減率は **m-LDA2** のほうが大きいものの統合後の精度は **m-LDA1** のほうがやや高く，異なる学習データセットを用いるメリットは確認できなかった．**m-LDA2** のように復元抽出による異なったデータセットから独立に複数の学習器を得る方式は bagging [22] として知られているが，他の方式，たとえば boosting [23] のように学習結果に基づいて逐次データセットを構成する方式をとれば違った結果となる可能性もある．また前述のように，モデル数をさらに増やした場合についても検証の余地がある．これらの点については今後の課題としたい．**m-LDA1** では各 LDA モデル間の差異は学習時の初期値の違いのみによるものである．今回の実験結果は，初期値の違いによって異なった学習結果に収束する不安定さをモデル統合により軽減できることを示しているともいえる．

（4）学習時の初期化方法との関係

LDA の学習時，パラメータ $\boldsymbol{\beta}$ の初期値は一様分布としパラメータ $\boldsymbol{\alpha}$ のみを乱数で初期化することで学習が可能である．しかし本論文の実験では **m-LDA1**，**m-LDA2** ともにパラメータ $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ 両方に乱数を与えて初期化した．またパラメータ $\boldsymbol{\alpha}$ については 1 以上の値を初期値として与えた．これらは初期化の自由度を高くすることによってモデル間の差異を拡大することを意図したものであるが，ここでは $\boldsymbol{\alpha}$ のみを乱数で初期化した場合の実験結果を示し，学習時の初期化方法と提案手法の効果との関係を考察する．

**図 13** に $\boldsymbol{\alpha}$ のみを乱数で初期化した場合の unigram PP を示す．ここでは $\sum \alpha_k = 1$ となるように $\boldsymbol{\alpha}$ を初期化した．初期化方法以外の実験条件は 4.2 節の **m-LDA1** と同一である．初期化の自由度が最小の場合であっても複数 LDA 統合により PP が減少し提案手法の効果が得られることが分かる．図 4 (a) と比較すると，やや意外な結果であるが統合による PP 削減効果は $\boldsymbol{\alpha}$ のみ乱数で初期化した場合のほうが大きい．単一 LDA での精度がやや低下しており，結果的に複数 LDA 統合後の PP は図 4 (a) とほぼ同程度となっている．

（5）最適な潜在トピック数およびモデル数

実装の際の最適な潜在トピック数 $C$ およびモデル数 $M$ の決定についてであるが，実行時の計算コスト・メモリコストはほぼ総トピック数 $C \times M$ で決まるため，実装可能な総トピック数の範囲内で最も高い精度が得られる構成とするのが望ましい．図 4 より $M \geqq 2$ の場合，$C \times M$ が同一であれば $C$ が大きいほど unigram の推定精度は高い．ただし $C = 50$ と $C = 70$ の差はわずかである．図 11 および図 12 からも $C = 50$ と $C = 70$ では PP の削減率はほぼ等しいことが分かり，$C$ をこれ以上増やしても統合モデルの精度向上はあまり期待できない．

したがって，基本的には単一 LDA でできるだけ高精度な潜在トピック数のモデルを，計算コストの許容範囲内で複数個組み合わせる．ただしモデル数 $M$ が小さい場合には（たとえば $M = 2$，3）統合による安定化の効果が十分得られない恐れがある（実際，図 5 (b) で

は $M = 2$ で $C = 30$，50，70 の場合に，$C \times M$ が等しく $C$ がより小さい構成のほうが高精度となっている）．このため，モデル統合の効果が十分得られるようモデル数 $M$ がおおむね 4〜5 以上となるようにすることが望ましい．

### 6.2　提案手法における改善事例の分析

5.2 節では提案手法で文脈依存 $N$-gram の推定精度が向上した記事の例を示したが，ここではこれらの例に着目し，提案手法で推定精度が向上するメカニズムについて考察する．

図 6 (d) の記事に対する単一 LDA での unigram PP は，潜在トピック数 $C = 50$ のモデル 1〜4 においてそれぞれ 3,548，2,480，1,460，2,352 であり，モデル 1 での精度が大きく劣っていた．一方，$C = 50$ のモデル 1〜4 統合時の PP は 1,828 であった．単一 LDA での PP が最も良いモデル 3 においてこの記事に対するトピック混合比（文脈適応時に求めた $\gamma_k$ を正規化した値）が最大値（= 0.942）となった潜在トピックについて，トピックの内容を表す特徴語を調べたところ[⋆1]，「大学」「学生」「試験」「教授」「科学」「研究」「大学院」「教育」「学部」「合格」といった単語が上位を占めた．

一方，モデル 1 で各潜在トピックの特徴語を調べたところ，教育・大学に関する内容のトピックは生成されていなかった．このモデルで図 6 (d) の記事に対し混合比が最大となったトピックの特徴語は「大阪」「京都」「地区」「野球」「大会」「商」「中国」「東京」「学園」「決勝」といった語であり，学生スポーツに関する内容のトピックであった．なお全評価記事に対するモデル 1〜4 の PP はそれぞれ 1,304，1,309，1,303，1,308 でありほぼ同等であった．

このように単一モデルでは，ある話題に関する潜在トピックが形成できなかった場合，似た話題を含む文書で精度低下が起こる．有限個の潜在トピック数でモデリングする以上，この現象を回避することは困難である．しかしモデルごとに苦手な話題は異なるため，複数のモデルの出力を統合することにより相互に補完し合い精度が向上する．これが本提案手法で精度が向上するメカニズムである．

### 6.3　先行研究との比較

LDA では各トピックがディリクレ分布に従って生成されると考えるのに対し，Dirichlet Mixture（DM）[7] は単語がトピックごとのディリクレ分布に従って生起すると考えるモデルである．DM は比較的低い混合数（潜在トピック数）で過適応を起こす傾向があるが，文献 7) では潜在トピック数の異なる複数の DM の出力を平均することにより過適応を抑制できることが示されている．

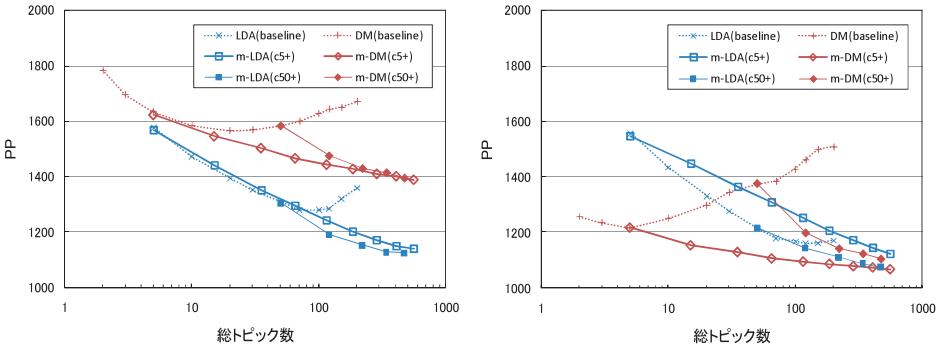



**図 14**　m-LDA と m-DM の比較
Fig. 14　Comparison between multiple-LDA and multiple-DM.

**図 14** に LDA と DM において複数モデルを統合した比較実験を行った結果を示す．学習テキスト・評価テキスト・語彙サイズ等の実験条件は前出の **m-LDA1** と同一である．DM のパラメータは leave-one-out 法[7] により推定した．PP が最小となる潜在トピック数が LDA と DM とで大きく異なり，同一トピック数の複数モデルを統合した場合の比較が困難なため，ここでは文献 7) と同様，異なるトピック数のモデルを組み合わせる方式で比較を行った．**m-LDA** および **m-DM** はそれぞれ複数 LDA および複数 DM の統合を表す．“(c5+)” は $C = 5$ のモデルから始め $C = 10$，20，30，50，70，100，120，150 のモデルを順に 1 つずつ追加していった場合を意味する．同様に “(c50+)” は $C = 50$ のモデルから始め $C = 70$，100，120，150 のモデルを順に追加した場合を示す．同一規模の単一モデル（baseline）との比較のため，**m-LDA**，**m-DM** とも総トピック数を横軸にとりプロットした．

図 14 (a) はここまでの実験同様，ヒストリ長 $L$ を固定しつねに直前の 20 形態素のみからトピック推定を行った場合であり，一方，同図 (b) はそこまでの全単語をヒストリとした場合である．(a)，(b) とも各記事の先頭 20 単語まではトピック推定を行わずトピック非依存 unigram 確率をそのまま用いた．図より，$L = 20$ に固定した場合 LDA が全般に優れている．これは DM では短いヒストリから安定してトピック推定を行うことが難しいためである．ただし “(c5+)” および “(c50+)” とも **m-DM** ではつねに baseline より PP を削減で

⋆1 トピックにおける unigram 確率 $p(w_j \mid z_k)$ とトピック非依存 unigram 確率 $p(w_j)$ との比が大きい語をトピックの特徴語とした．

きており，複数モデル統合により性能が向上している．**m-LDA** も総トピック数の増加にともなって PP が減少するが，“(c5+)” では総トピック数 = 65（すなわち 5，10，20，30 トピックの 4 モデル統合）までは baseline と同程度の性能にとどまる．“(c50+)” ではつねに baseline を上回るため，ある程度高性能なモデルどうしを組み合わせる必要がある．一方，$L = 20 \sim \infty$ の場合には DM の精度が大幅に向上し，少ないトピック数で **m-DM** の性能が LDA および **m-LDA** を上回る．**m-LDA** はここでも “(c50+)” が “(c5+)” より良い結果となった．

以上より，短いヒストリから最も精度良く推定できるのは **m-LDA** である．我々のターゲットであるテキスト入力支援，特に医療現場でのテキスト入力支援では，診察する患者が変われば話題（疾患群・疾患部位等）も変化し，同じ話題が続くのはせいぜい数十単語と考えられる．したがってこのような場面では提案手法が最も適している．ただし他の応用，たとえば文書分類等文書全体からトピック推定を行う場合や，入力支援であっても話題変化がきわめて緩やかなケース等では，**m-DM** のほうがよりコンパクトなモデルで高い性能を得られる可能性がある．DM では適応時に反復計算が不要なため，低計算コストが要求される場面にも有効である．

なお，LDA に関しては HDP（階層ディリクレ過程）を用いて最適な潜在トピック数を推定できることが報告されている[14]．ただし HDP の性能は単一 LDA の最高性能と同等[14]なため本提案手法と HDP との直接の比較は行っていない．HDP で推定したトピック数が本提案手法で複数 LDA を統合する際の個々のモデルのトピック数決定に利用できる可能性があると考えており，今後，検証を行う予定である．また DM のパラメータ推定に関しては，階層ベイズモデルを用いて平滑化を行い過適応を改善できること（smoothed-DM）が報告されているが[13]，この方式との比較は今後の課題としたい．

### 6.4　テキスト入力支援における有効性

5.4 節では，提案方式のトピックモデルをテキスト入力支援に適用することにより，トピック非依存の $N$-gram モデルよりも予測候補数を削減できることを示した．ここではテキスト入力支援の必要性を概説するとともに，本論文で提案した評価指標と言語モデルのテキスト入力支援における有効性を考察する．

(1) テキスト入力支援の必要性

予測入力に基づくテキスト入力支援は，入力負荷を軽減する技術として特にキーの数が少ない携帯電話等で広く利用されている[24]．重度身障者の言語入力を支援するユニバーサル技術として予測入力を用いた研究も報告されている[25]．筆者らの一部は予測入力によるテキスト入力支援が電子カルテシステムの入力効率向上に有効であることを確認しており[12),26)]，実際に製品に搭載され利用されている．

電子カルテシステムは通常のパーソナルコンピュータ同様，フルキーボードによりテキストを入力するが，医療現場では限られた時間内に多くの患者を診察する必要があり，できる限り迅速に入力したいというニーズがきわめて高い．従来は汎用のかな漢字変換システムに医療用語辞書を追加して利用する形態が一般的であったが，病名・薬品名等の専門用語には長い単語も多いため入力の負荷を軽減できる予測入力が注目されている．トピックモデルを入力支援に適用した例は筆者らの知る限りまだなく，トピック適応により現行の入力支援システムをさらに改良できればその意義は大きい．

(2) 評価指標 i-PP の意義

PP はテキスト入力支援においても言語モデルの基本的な評価指標として有用である．しかし PP を削減できたとしてそれが実際の応用においてどの程度のメリットをもたらすかは必ずしも明確でない．テキスト入力支援で重要なのは，読みを 1 文字ずつ入力していったときに言語モデルによって上位何位までに候補を絞り込むことができるかであるが，PP から直接これを知ることはできない．

入力支援システムの入力効率を測る指標として，筆者らは以前，入力読み文字数とキータッチ数を用いた[12]．これらの指標は，ユーザに提示する候補単語数 $k$ を決めておき，読みを 1 文字ずつ入力し正解候補が上位 $k$ 位以内に現れた時点で選択するものとして算出する．しかしこれらの値は候補単語数 $k$ に依存するため，システム全体の性能の目安にはなるが言語モデル自体の評価指標とはいい難い．

これに対し本論文で提案した i-PP は任意の入力読み文字数における平均単語分岐数を表す．読みを 1 文字ずつ入力していったときに上位何位までに正解候補を絞り込むことができるかの目安を知ることができ，入力支援システムにおける言語モデル自体の能力を測る指標となる．5.4 節では，**m-LDA** を入力支援に適用した場合，入力読み文字数 $l = 1 \sim 2$ では $l = 0$ のとき（通常の PP）と同程度パープレキシティを削減できることを示した．これは i-PP が PP の拡張となっているために得ることのできた結果である．

(3) テキスト入力支援における提案方式の有効性

本論文の実験では，入力読み文字数 $l = 2$ のときトピック非依存 $N$-gram と比較して i-PP を bigram で 23.3%，trigram で 17.9%削減でき，i-PP を 5 未満に抑えることができた．ユーザに提示する候補数は 5 語程度が望ましいため，この結果は **m-LDA** を入力支援に適用することにより読みを 2 文字入力すれば候補リスト内に正解候補が含まれると期待でき

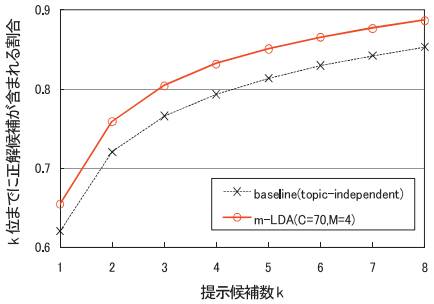



**図 15**　正解候補が $k$ 位までに含まれる割合（$l = 2$）
Fig. 15　The rate at which the correct candidate is contained among the $k$-best candidates ($l = 2$).

ることを意味している．

**図 15** は，5.4 節の実験結果から入力読み文字数 $l = 2$ の時点において候補リスト内に正解候補が含まれる確率がどの程度であるかを調べた結果である（trigram の場合）．横軸は提示する候補単語数 $k$ であり，縦軸は上位 $k$ 位までに正解候補が含まれる割合を表している．トピック非依存 trigram では上位 5 位までに正解が含まれる割合は 0.813 であるが，**m-LDA**（$C = 70$，$M = 4$）では 0.851 に向上している．これは一見，小さな変化のように思えるが，トピック非依存 trigram で同程度の割合で正解候補が含まれる候補数は 8 である．すなわち提案方式で候補数を 5 とした場合，トピック非依存 trigram で候補数を 8 とした場合と同程度の効果を提供することができ，候補リストの一覧性を大幅に改善することができる．

**図 16** に本提案手法で大幅な改善が見られた例を示す．コンテキストは図 6 (d) の記事の一部であり，正解語「留学生」を予測した際の 10 位までの候補リストをあわせて示した．「モデル 1 単独」は $C = 50$ の単一 LDA，「モデル 1〜4 統合」は $C = 50$ の LDA を 4 モデル統合した場合，「トピック非依存」はトピック非依存 trigram による結果であり，括弧内の数字はそれぞれの方式による trigram 確率である．「モデル 1 単独」では正解語の順位が「トピック非依存」よりも落ちているが，「モデル 1〜4 統合」では 1 位となっている．単一 LDA でトピック推定に失敗したのを他のモデルで補完できた一例である．

## 7. おわりに

複数個の LDA トピックモデルを統合する方式をテキスト入力支援に適用・高精度化することを目的とし，提案方式を用いて $N$-gram をトピックに適応させた場合の推定精度と安定性を検証した．さらにテキスト入力支援に適した言語モデル評価指標 i-PP を提案し，この指標を用いた評価を行った．

実験の結果，提案方式では過適応による性能低下が抑制され同規模の単一 LDA よりつねに性能が向上すること，統合により推定精度が安定化することが確かめられた．特に trigram をトピック適応させた場合，単一 LDA に比べ大幅に PP を削減できる．学習データセットと統合効果との関係に関しては，提案方式では異なる学習データで学習したモデルを統合した場合に統合の効果が向上すること，学習データ量が少ない場合でも従来より高い精度を実現できることを確認した．

本論文で提案した評価指標 i-PP は任意の入力読み文字数における平均単語分岐数を表し，入力支援システムにおける言語モデルの能力を測る指標となる．この指標を用いて評価を行った結果，提案方式では入力読み文字数 $l = 2$ まで通常の PP と同程度にパープレキシティを削減できることが確認できた．また，入力読み文字数 $l = 2$ のとき i-PP を 5 未満に抑えることができ，従来方式よりも高精度に正解候補を絞り込めることが確かめられた．ただし今回の実験は新聞記事コーパスを用いて基本的な振舞いを調べた段階である．ターゲットである医療分野をはじめ，様々な分野のテキストを対象とした検証は今後の課題の 1 つ

［コンテキスト］
…試験は来月８日以降。募集定員は一般入試分を掲載し、
外国人*留学生*

| ［モデル 1 単独］ | ［モデル 1〜4 統合］ | ［トピック非依存］ |
|---|---|---|
| の (0.2064) | 留学生 (0.2092) | の (0.1428) |
| 女性 (0.0878) | の (0.1623) | 投資家 (0.0980) |
| が (0.0734) | 講師 (0.0405) | を (0.0549) |
| を (0.0499) | を (0.0395) | が (0.0538) |
| 男性 (0.0498) | が (0.0379) | 観光客 (0.0538) |
| 登録 (0.0457) | 力士 (0.0374) | 労働者 (0.0448) |
| は (0.0333) | 旅行 (0.0285) | 留学生 (0.0414) |
| に (0.0308) | 女性 (0.0237) | に (0.0300) |
| 旅行 (0.0232) | 登録 (0.0236) | 選手 (0.0285) |
| 労働者 (0.0182) | に (0.0229) | 登録 (0.0254) |
| ： | | |

「留学生」は 25 位

**図 16**　コンテキストと予測候補の例
Fig. 16　An example of a context and candidate lists.

である．

本論文では $N$-gram をトピック適応させる際，カッツ・スムージングにより平滑化したトピック非依存 $N$-gram を用いたが，最近，従来よりも高精度な $N$-gram を構築する手法が提案されている[27),28)]．これらの技術を用いることにより本提案方式をさらに改善できる可能性がある．また学習条件と LDA 統合効果との関係に関しては，統合モデル数をさらに増やした場合について検証を行うとともに，異なる学習データセットにより学習を行う方式についても再検討する必要があると考えている．さらに他のトピックモデルとの比較，特に smoothed-DM およびこれを複数個統合した場合との比較実験にも取り組む予定である．

## 参 考 文 献


1) Jelinek, F.: Self-organized Language Modeling for Speech Recognition, *Readings in Speech Recognition*, pp.450–506, Morgan Kaufmann Publishers (1990).
2) Kuhn, R. and de Mori, R.: A Cache-based Natural Language Model for Speech Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.12, No.6, pp.570–583 (1990).
3) Tillmann, C. and Ney, H.: Word Triggers and the EM Algorithm, *Proc. CoNLL-97*, pp.117–124 (1997).
4) Thrun, S., Nigam, K., McCallum, A. and Mitchell, T.: Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, Vol.39, No.2/3, pp.103–134 (2000).
5) Hofmann, T: Probabilistic latent semantic indexing, *Proc. 22nd Annual ACM Conference on Research and development in Information Retrieval*, pp.50–57 (1999).
6) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
7) 貞光九月，三品拓也，山本幹雄：混合ディリクレ分布を用いたトピックに基づく言語モデル，電子情報通信学会論文誌 D-II，Vol.J88-D-II, No.9, pp.1771–1779 (2005).
8) Gildea, D. and Hofmann, T.: Topic-based Language Models using EM, *Proc. Eurospeech'99*, pp.2167–2170 (1999).
9) 高橋力矢，峯松信明，広瀬啓吉：複数のバックオフ N-gram を動的補間する言語モデルの高精度化，情報処理学会研究報告 SLP-49-11, pp.61–66 (2003).
10) 根本雄介，秋田祐哉，河原達也：講義音声認識のためのスライド情報を用いた言語モデル適応，言語処理学会第 13 回年次大会論文集，pp.131–134 (2007).
11) 三品拓也，貞光九月，山本幹雄：確率的 LSA を用いた日本語同音異義語誤りの検出・訂正，情報処理学会論文誌，Vol.45, No.9, pp.2168–2176 (2004).
12) 中村　明，川尻博光，金川　誠，松本忠博，池田尚志，速水　悟，紀ノ定保臣：統計的言語モデルに基づく電子カルテ入力支援システムの開発，言語処理学会第 13 回年次大会論文集，pp.998–1001 (2007).
13) 貞光九月，待鳥裕介，山本幹雄：混合ディリクレ分布パラメータの階層ベイズモデルを用いたスムージング法，情報処理学会研究報告 SLP53-1, pp.1–6 (2004).
14) Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M: Sharing Clusters among Related Groups: Hierarchical Dirichlet Process, *NIPS 2004*, MIT Press (2004).
15) 中村　明，津田裕亮，松本忠博，池田尚志，速水　悟：複数モデルの統合による LDA トピックモデルの高精度化，言語処理学会第 14 回年次大会論文集，pp.305–308 (2008).
16) Griffiths, T.L. and Steyvers, M.: Finding Scientific Topics, *PNAS*, Vol.101, pp.5228–5235 (2004).
17) 毎日新聞社：CD-毎日新聞 2005/2006 データ集本社版，日外アソシエーツ (2006/2007).
18) 山田佳裕，脇田貴之，大口智也，池田尚志：文節構造解析システム ibukiC の解析仕様および精度の比較と評価，言語処理学会第 13 回年次大会論文集，pp.167–170 (2007).
19) Minka, T.: Estimating a Dirichlet Distribution (2003). http://www.stat.cmu.edu/~minka/papers/dirichlet/
20) Katz, S.M.: Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol.35, No.3, pp.400–401 (1987).
21) Miller, G.A.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, *The Psychological Review*, Vol.63, pp.81–97 (1956).
22) Breiman. L: Bagging Predictors, Technical Report 421, Statistics Department, Univ. of California, Berkeley (1994).
23) Schapire, R.: The Strength of Weak Learnability, *Machine Learning*, Vol.5, pp.197–227 (1990).
24) 増井俊之：携帯端末のテキスト入力手法，ヒューマンインタフェース学会誌，Vol.4, No.3, pp.131–144 (2002).
25) 田中久美子：重度身障者のための 1 ボタン自然言語入力システム，言語処理学会第 10 回年次大会論文集，pp.544–547 (2004).
26) 川尻博光，中村　明，金川　誠，松本忠博，池田尚志，速水　悟，紀ノ定保臣：予測入力における医療用言語モデルの有効性評価，第 27 回医療情報学連合大会論文集，pp.458–461 (2007).
27) Teh, Y.W.: A Hierarchical Bayesian Language Model based on Pitman-Yor Processes, *Proc. COLING/ACL 2006*, pp.985–992 (2006).
28) 持橋大地，隅田英一郎：階層 Pitman-Yor 過程に基づく可変長 n-gram 言語モデル，情報処理学会論文誌，Vol.48, No.12, pp.4023–4032 (2007).

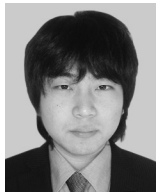

(平成 20 年 7 月 10 日受付)
(平成 21 年 1 月 7 日採録)

**中村　　明**（正会員）

1989 年名古屋大学工学部電気系学科卒業．1991 年同大学大学院工学研究科電気電子工学専攻修士課程修了．同年三洋電機株式会社入社．以来，手書き文字認識，自然言語処理，医療情報処理の研究に従事．岐阜大学大学院工学研究科電子情報システム工学専攻博士後期課程在籍中．電子情報通信学会会員．

**速水　　悟**（正会員）

1981 年東京大学大学院工学系研究科修士課程修了．同年通商産業省工業技術院電子技術総合研究所入所，2001 年独立行政法人産業技術総合研究所，2002 年より岐阜大学工学部応用情報学科教授．この間，1989〜1990 年米国カーネギーメロン大学客員研究員，1994〜1995 年フランス国立情報機械研究所客員研究員．博士（工学）．音声情報処理，マルチモーダル情報処理に関する研究に従事．日本音響学会，電子情報通信学会，人工知能学会，IEEE，ACM 各会員．



**津田　裕亮**

2007 年岐阜大学工学部応用情報学科卒業．2009 年同大学大学院工学研究科応用情報学専攻修士課程修了（見込）．自然言語処理の研究に従事．

**松本　忠博**（正会員）

1985 年岐阜大学工学部電子工学科卒業．1987 年同大学大学院工学研究科電子工学専攻修士課程修了．博士（工学）．現在，岐阜大学工学部応用情報学科助教．自然言語処理，手話言語工学の研究に従事．電子情報通信学会，言語処理学会，日本ソフトウェア科学会，日本手話学会各会員．

**池田　尚志**（正会員）

1968 年東京大学教養学部基礎科学科卒業．同年工業技術院電子技術総合研究所入所．制御部情報制御研究室，知能情報部自然言語研究室に所属．1991 年岐阜大学工学部電子情報工学科教授．現在，同応用情報学科教授．工学博士．自然言語処理，人工知能の研究に従事．電子情報通信学会，人工知能学会，言語処理学会各会員．